Argonne NATIONAL LABORATORY

Concerted Flows: Infrastructure for Terabit/s Data Transfer

Raj Kettimuthu, Venkat Vishwanath, Ian Foster, Bob Grossman, Mark Hereld, Mike Papka and Steve Tuecke

U.S. DEPARTMENT OF ENERGY

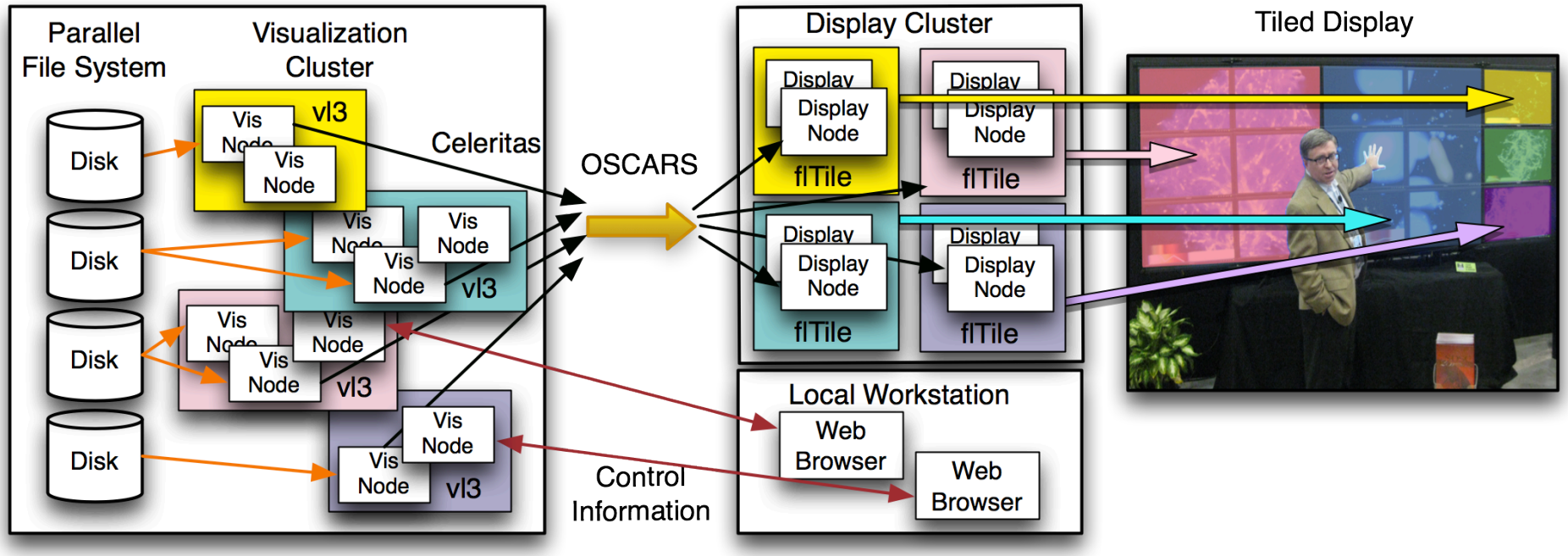
Outline

- Overview
- Motivating applications
- Data movement characteristics
- Network characteristics and End system trends
- Concerted Flows API design
- Preliminary results

Concerted Flows

- **Problem**: Traditional data transfer protocols fail to scale to high-speed networks as well as end-systems, and do not effectively satisfy the diverse needs of applications
- **Innovations**: Develop new parallel protocols that are
  - **Composable:** Captures the diverse flow characteristics and needs
  - **Adaptive:** Leverages feedback from network agents and exploits topology to design flow and congestion control for parallel data movement
- **Impact**: Building a knowledge base capturing the data transfer patterns of several DOE applications.
- Design of an API and framework for parallel data movement that caters to characteristics of applications, future architecture and shared infrastructure

Interactive Remote Visualization of ENZO Cosmology



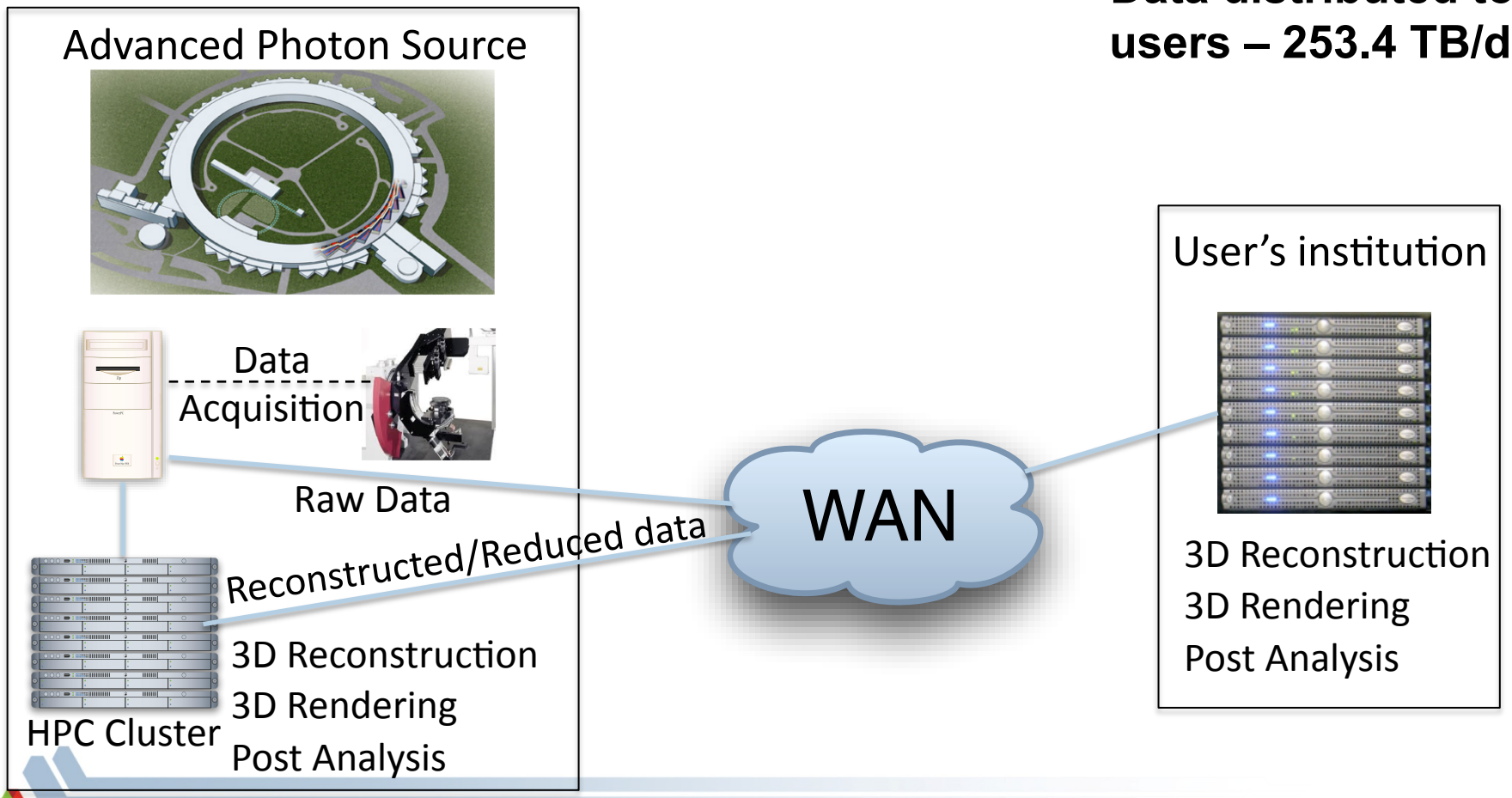
Argonne National Laboratory

San Diego

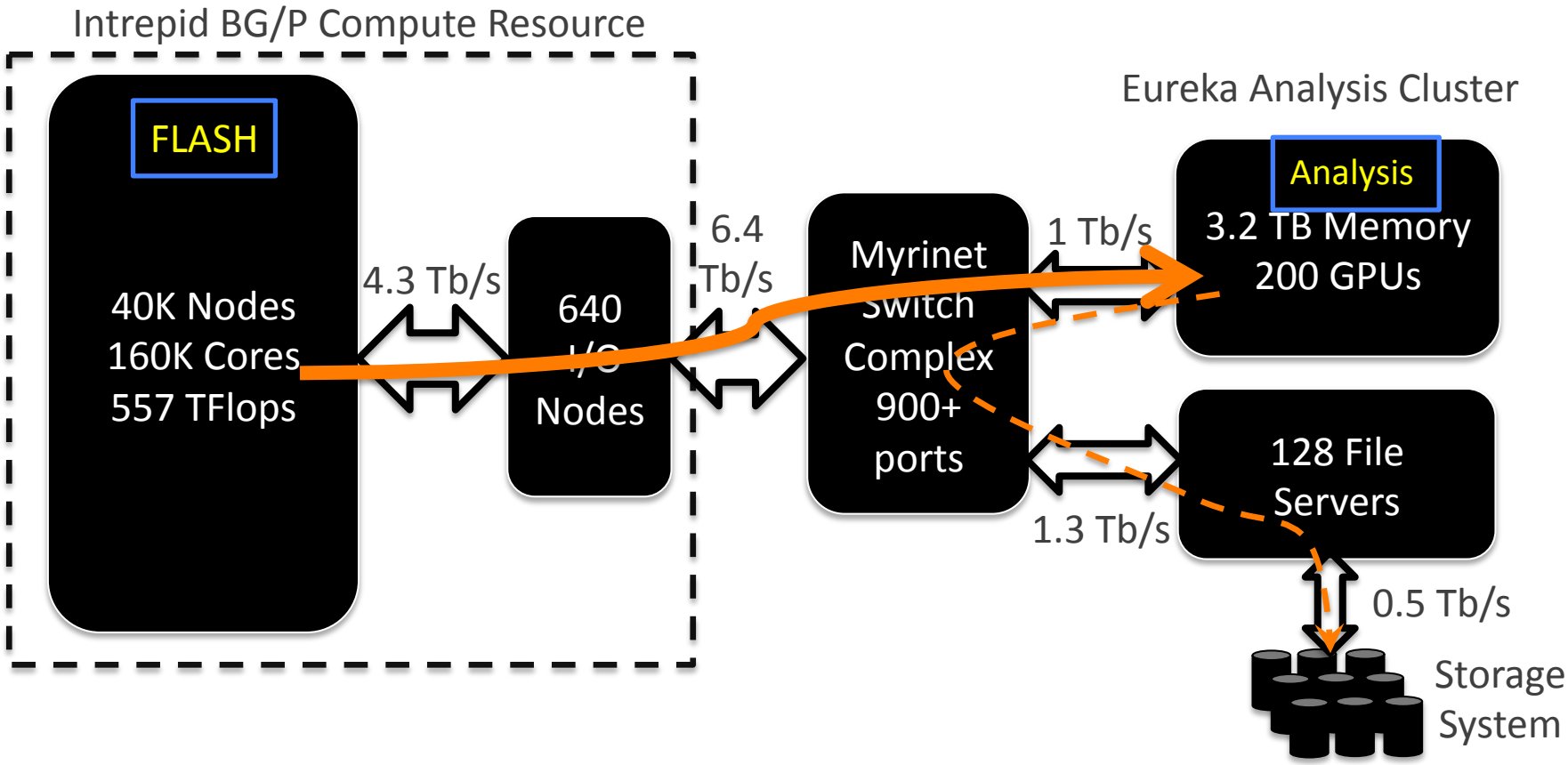
New Orleans - SC'10 Show floor

Tomography at APS

- **Current**
  - **Data processed – 5.6 TB/day**
  - **Data distributed to users – 3.3 TB/day**
- **Upgrade**
  - **Data processed – 385.3 TB/day**
  - **Data distributed to users – 253.4 TB/day**

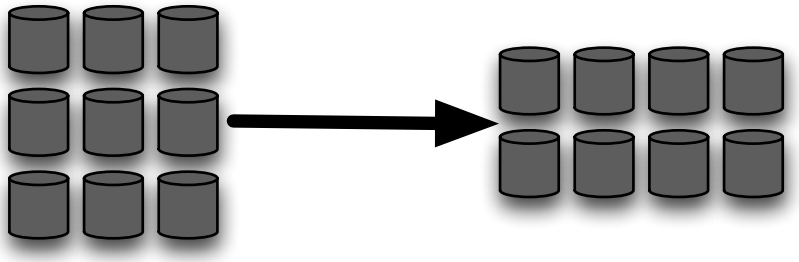


Simulation-time Data Analysis and Visualization of FLASH Astrophysics Simulation

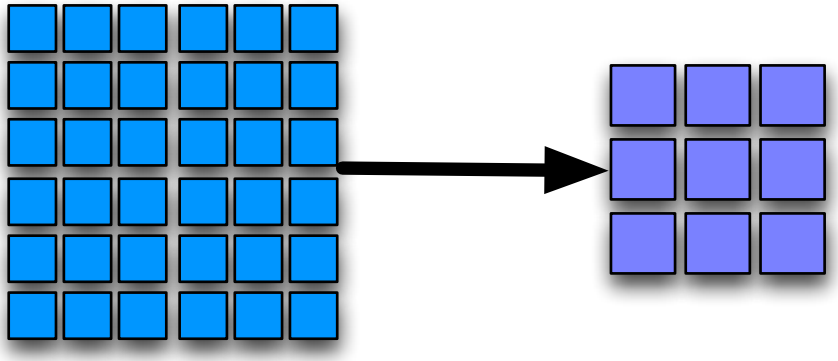


Simulation-time data analysis is critical to reduce the data written to storage and to generate faster insights

Data Movement Trends

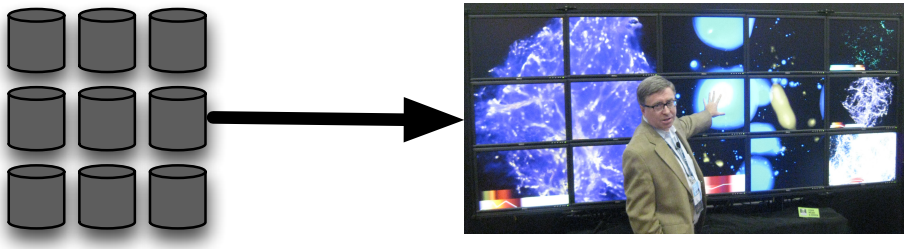


Disk-to-Disk Transfers

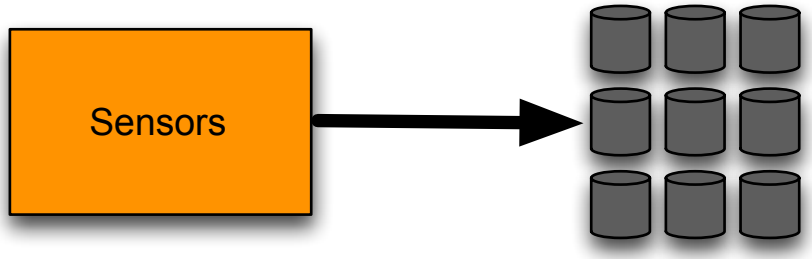


Memory-to-Memory Transfers

Parallel M-to-N Data Flows in a shared infrastructure



Disk-to-Memory Transfers



Memory-to-Disk Transfers

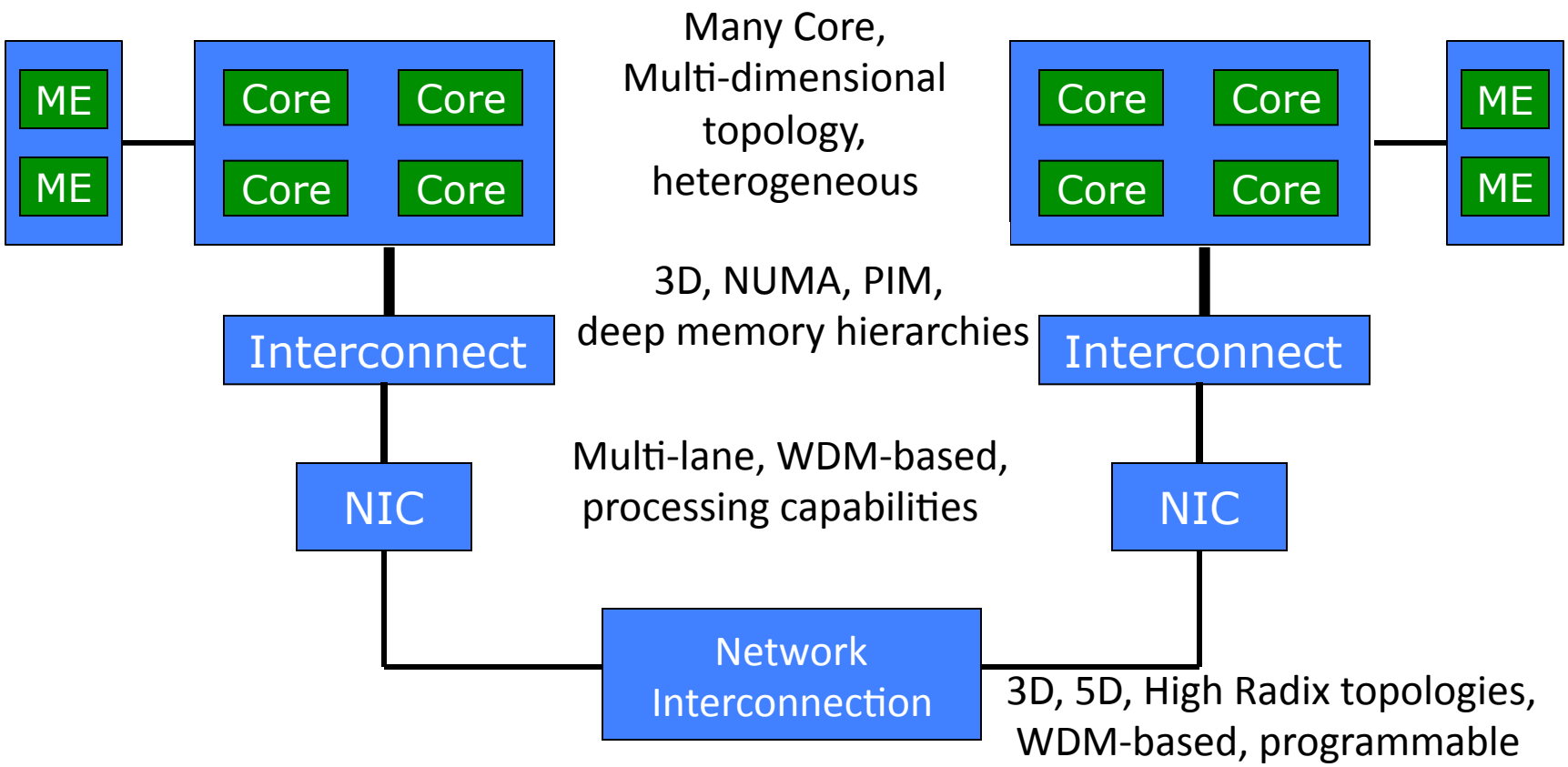
Characteristics of Application Flows

App	Type of Flow	# of Flows	BW	Latency	Burstiness	Size	Protocol
Globus Online	Data	n per node	High	N	Y	Variable	TCP, UDT
	Control	1 per session	Low	Y	Y	Small	TCP
APS	Data	n per detector	High	N*	Y	Large	TCP
	Control	1 per app	Low	Y	Y	Small	TCP
FLASH Simulation-time Analysis	Data	1 per core	High	N*	Y	Variable	TCP, RDMA
	Control	1 per app	Low	Y	y	Small	TCP, RDMA
ENZO Remote Viz	Data	1 per display	High	Y	N	Large	TCP, UDP
	Control	1 per app	Low	Y	Y	Small	TCP

Network Characteristics

- Network Type
  - Shared or dedicated
  - Circuit or packet or hybrid
- Network activity
  - Over-utilized or under-utilized
- Network Topology
  - Parallel paths
  - Bandwidth, latency, loss rate
- LAN (within a leadership facility), MAN or WAN
- Network is no longer a blackbox and one can obtain monitoring information (perfSONAR) as well as provision/ configure current and future networks (OSCARS)

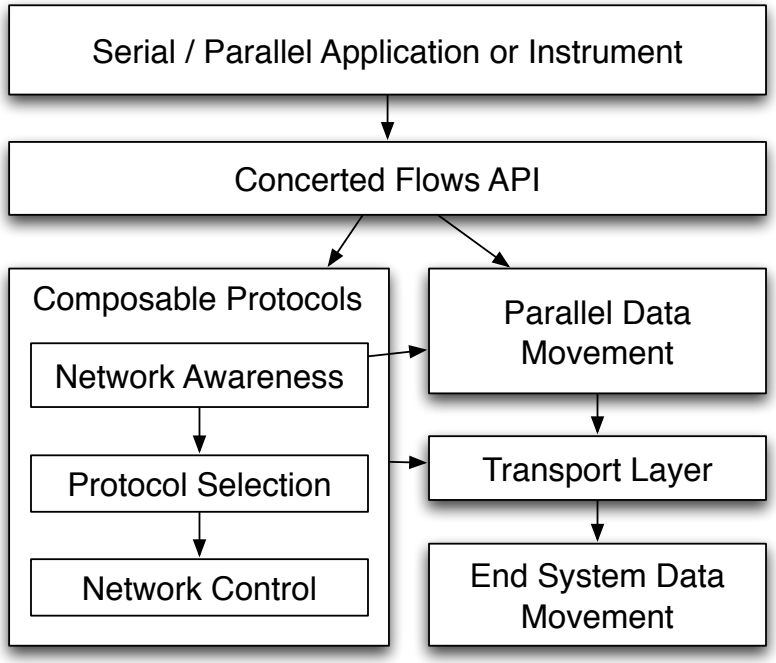
System Trends



Applications need to contend with the deep and complex system hierarchies and take advantage of parallelism in the various sub-systems

Objectives

- Develop concerted flows API
  - Capture the requirements of the application
  - Capture the characteristics of various components in the end-to-end path
    - Network, End-systems
- Create data transfer benchmark kernels for representative applications
  - Flash, Enzo, Select APS beamlines, Globus Online

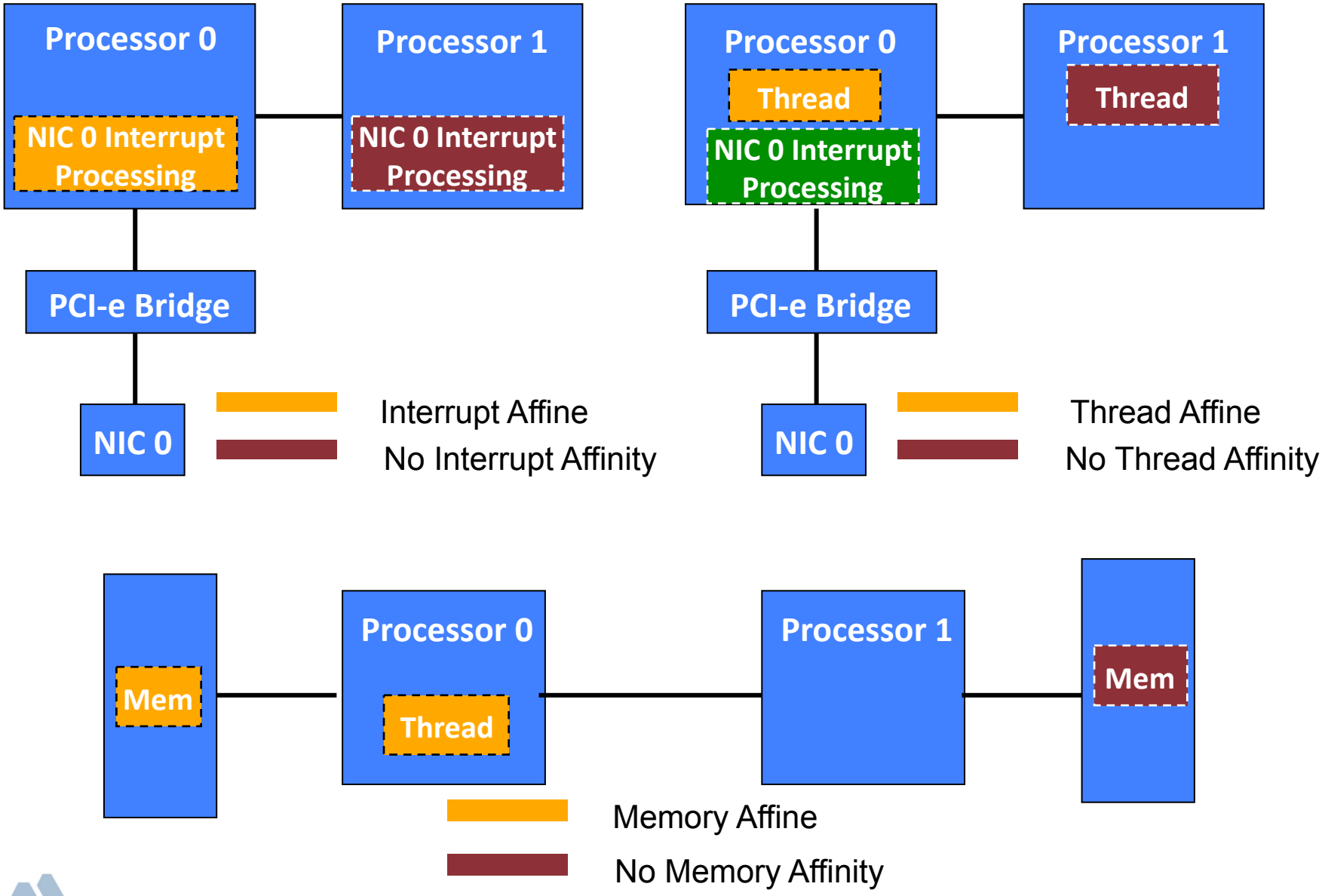


- Develop the concerted flows framework and focus primarily on end-systems and local area network
- Develop a parallel M-to-N data movement benchmark for concerted flows

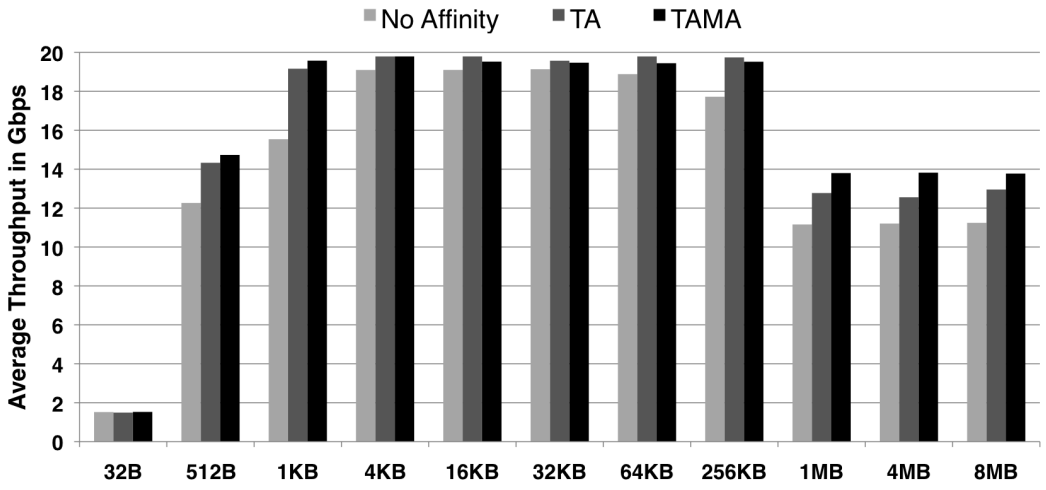
Concerted Flows API

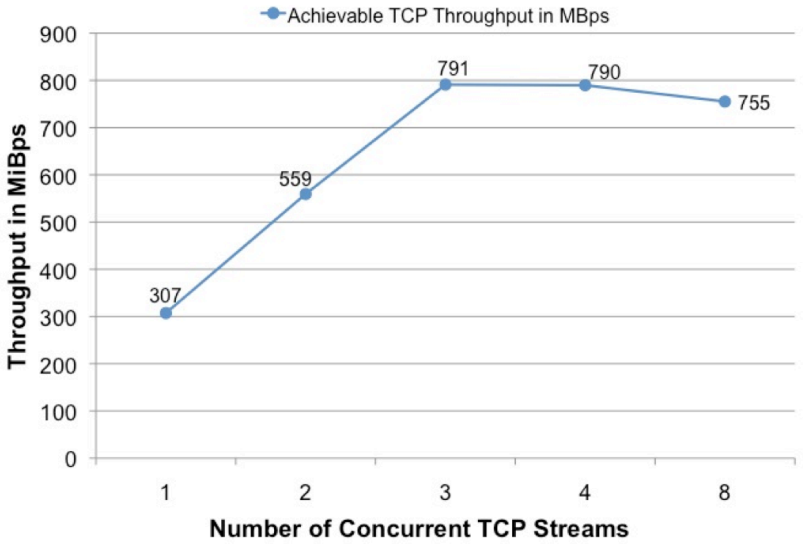
- Application requirements
  - M-to-N flows
  - Latency, Jitter, Bandwidth
  - Reliability, error rate
  - Burstiness, Deadline, Start time
  - Contiguous or non-contiguous
- Network characteristics
  - Loss rate, latency, bandwidth, QoS
  - Topology (parallel links, circuits, intermediate nodes)
- End system characteristics
  - Cores, memory, NIC
  - Topology Information
  - Storage (Available disk space, File system, optimal block size)

System Affinities



End System Topology-aware Data Movement

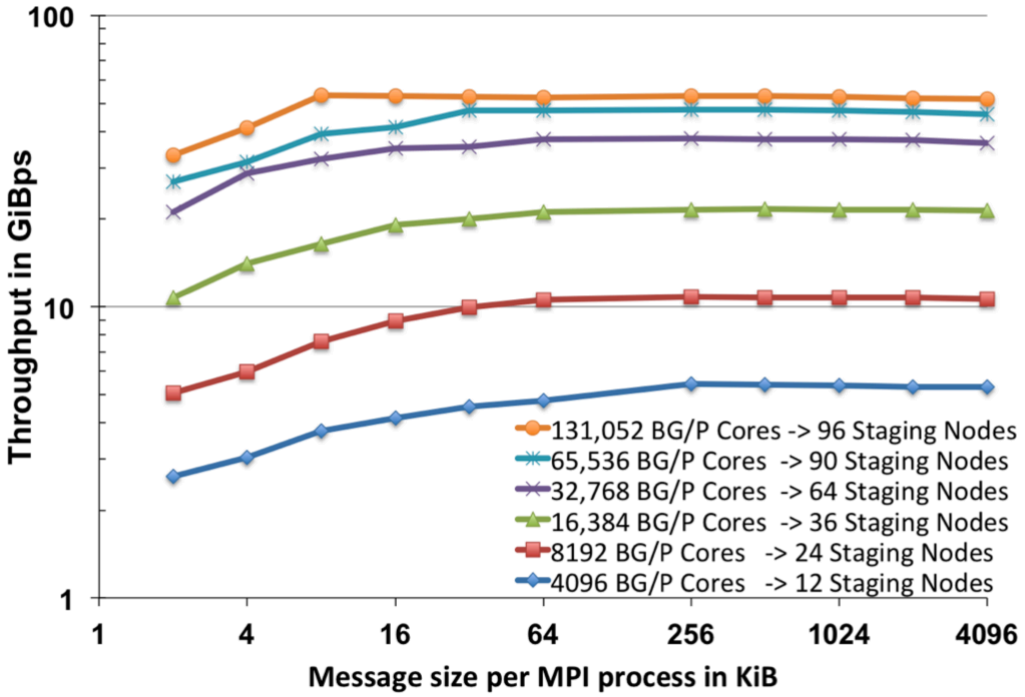




- System Affinities play a key role to achieve higher throughput.
- 28% improvement in throughput (from 17.2 Gbps to 22 Gbps) for RDMA based data transfer between two nodes connected by QDR Infiniband

- As we move towards future systems with many low-power cores, we need to leverage **parallel communication for improved performance**

Parallel Data Movement

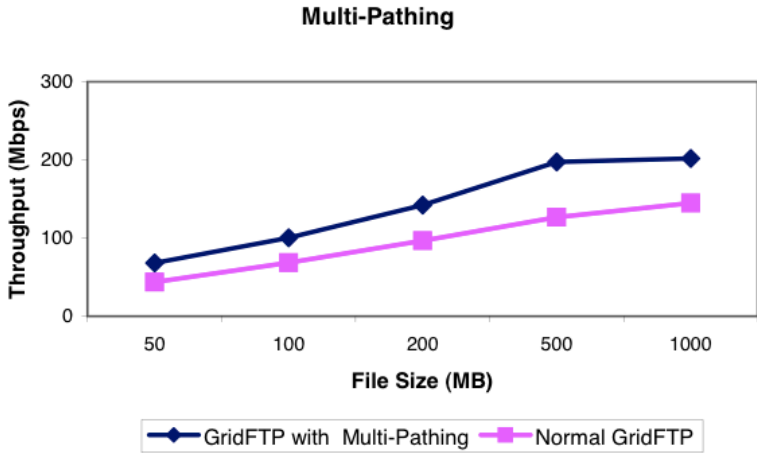


Data movement performance from ALCF Intrepid BG/P supercomputer to Data Analysis and Visualization cluster over Local Area Network

Data movement performance from OSU to Japan when multiple independent paths are utilized

OSU->ORNL->Japan

OSU->Starlight->Japan



Team

- Raj Kettimuthu
- Venkat Vishwanath
- Ian Foster
- Bob Grossman
- Mark Hereld
- Steve Tuecke
- Jun Yi (postdoc hire)

- http://wiki.mcs.anl.gov/concerted-flows/index.php/Main_Page